Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Journal Track

Confidence-based Estimators for Predictive Performance in Model Monitoring (Abstract Reprint)

Juhani Kivimäki[1] , **Jukka K. Nurminen**[1] , **Jakub Białek**[2] , **Wojtek Kuberski**[2]
[1]University of Helsinki, Finland
[2]NannyML
juhani.kivimaki@helsinki.fi, jukka.k.nurminen@helsinki.fi, jakub@nannyml.com, wojtek@nannyml.com

Abstract Reprint. This is an abstract reprint of a journal article by [Kivimäki *et al.*, 2025].

Abstract

After a machine learning model has been deployed into production, its predictive performance needs to be monitored. Ideally, such monitoring can be carried out by comparing the model's predictions against ground truth labels. For this to be possible, the ground truth labels must be available relatively soon after inference. However, there are many use cases where ground truth labels are available only after a significant delay, or in the worst case, not at all. In such cases, directly monitoring the model's predictive performance is impossible.

Recently, novel methods for estimating the predictive performance of a model when ground truth is unavailable have been developed. Many of these methods leverage model confidence or other uncertainty estimates and are experimentally compared against a naive baseline method, namely Average Confidence (AC), which estimates model accuracy as the average of confidence scores for a given set of predictions. However, until now the theoretical properties of the AC method have not been properly explored. In this paper, we bridge this gap by reviewing the AC method and show that under certain general assumptions, it is an unbiased and consistent estimator of model accuracy. We also augment the AC method by deriving valid confidence intervals for the estimates it produces. These contributions elevate AC from an ad-hoc estimator to a principled one, encouraging its use in practice.

We complement our theoretical results with empirical experiments, comparing AC against more complex estimators in a monitoring setting under covariate shift. We conduct our experiments using synthetic datasets, which allow for full control over the nature of the shift. Our experiments with binary classifiers show that the AC method is able to beat other estimators in many cases. However, the comparative quality of the different estimators is found to be heavily case-dependent.

References

[Kivimäki *et al.*, 2025] Juhani Kivimäki, Jukka K. Nurminen, Jakub Białek, and Wojtek Kuberski. Confidence-based estimators for predictive performance in model monitoring. *J. Artif. Int. Res.*, 82, April 2025.